

CSC411 Machine Learning

Project 3: Fake News

Ariel Kelman

Student No: 1000561368

Gideon Blinick

Student No: 999763000

14 March 2018

1 Dataset Description

Both datasets (real and fake) contain headlines about U.S. President Donald Trump. All characters are lowercase. It is certainly feasible to determine whether a headline is real or fake news based on the presence or absence of certain keywords in the headlines. For the part, we analyzed the data in 2 ways:

- First we looked at the words that had the largest difference in percentage for appearance in the datasets. For example, the words "the", "trump" and "hillary" appeared more often in the fake news headlines by percentage by 20.0%, 10.4% and 10.3%, respectively, while the words "donald", "trumps", and "us" appeared more often in the real news headlines by 24.5%, 10.8%, and 8.8%, respectively.
- Second, we looked at words that only appear in one dataset and not the other by count. Count was necessary here, because percentage difference cannot apply. So the top words in the fake dataset that did not appear in the real dataset were "breaking" (27 headlines), "u" (19), and "soros" (18), while the top words in the real dataset that were not in the fake one were "korea" (79), "turnbull" (55), and "travel" (52). These words are especially relevant because their presence in a headline allows us to automatically classify the headline as fake or real.

- 2 Naive Bayes Implementation
- 3 Predictive Factors
- 4 Logistic Regression
- 5 Logistic Regression vs. Naive Bayes
- 6 Miscle
- 7 Decision Tree
- 8 Decision Tree - Information Theory